Probabilistic Anomaly Detection in Dynamic Systems (Summary)

Padhraic Smyth
Jet Propulsion]] Laboratory 238-420
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
email: *pjs@galway.jpl.nasa .gov*
tel: (818) 3543768, fax: (818) 3546825.

Poster: Applications

1 Introduction

This paper describes probabilistic methods for novelty detection when using pattern recognition methods for fault monitoring of dynamic systems. 'l'he problem of novelty detection is particularly acute when prior knowledge and data only allow one to construct an incomplete prior model of the system. Hence, some allowance must be made in model design so that a classifier will be robust to data generated by classes not included in the training phase. For the fault detection problem a practical approach is to construct both an input density model and a discriminative class model. The construction of an input model for data of unknown origin is fundamentally ill-posed but can be solved in practical terms by using known constraints on the input features and constructing a non-informative prior density. In conjunction with Bayes rule, and some prior estimates of the relative likelihood of data of known and unknown origin, the actual classification equations are straightforward. The paper describes the application of this method in the context of hidden Markov models for online fault monitoring of large ground antennas for spacecraft tracking, with particular application to the detection of transient behaviour of unknown origin.

2 Problem Background

Conventional control-theoretic models for fault detection rely on an accurate model of tile plant being monitored: 'frequently in practice no such model exists for complicated non-linear systems. The large ground antennas used by JPL's Deep Space Network (DSN) to track planetary spacecraft fall into this category — quite complicated analytical models exist for the electro-mechanical pointing systems, but they are known to be a poor fit for fault detection purposes.

2.1 Basic Detection Architecture

We have previously described the application of online adaptive pattern recognition methods to this problem [1, 2]. The system operates as follows, Sensor data such as motor current, position encoder, tachometer voltages, and so forth are synchronously sampled at 50Hz by a data acquisition system. The data is blocked off into disjoint windows ('200 samples are used in practice) and various features (such as estimated autoregressive coefficients) are extracted; let the feature vector be $\underline{\theta}$.

The features are feel into a classification model (every 4 seconds) which in turn provides posterior probability estimates of the m possible states of the system given the estimated features from that window, $p(\omega_i|\underline{\theta})$. ω_1 corresponds to normal conditions, the other ω_i's, $1 \leq i \leq$ m, correspond to known fault conditions.

Finally, since the system has "memory" in the sense that it is more likely to remain in the current state than to change states, the posterior probabilities need to be correlated over time. This is acheived by a standard first-order hidden Markov model (HMM) which models the temporal state dependence [2].

As described in [1, 2] the classifier portion of the model is trained using simulated hardware "faults. The feed-forward neural network has been the model of choice for this application because of its discrimination ability, its posterior probability estimation properties [3, 4] and its relatively simple implementation in software. Also described in ['2] at length is the design of the HMM transition matrix based on prior knowledge of system mean time between failure (MTBF) information and other specific knowledge of the system configuration.

3 limitations of the Discriminative-HMM Model

The model described above assumes that there are m known mutually exclusive and exhaustive states (or "classes") of the system, ω_1 ,... ,ω_m. The mutually exclusive assumption is reasonable in many applications where multiple simultaneous failures arc highly unlikely. However, the exhaustive assumption is somewhat impractical. In particular, for fault detection in a complex system such as the antenna, there are literally thousands of possible fault conditions which might occur. The probability of occurrence of any single condition is very small, but nonetheless there is a significant probability that at least one of these conditions will occur over some finite time. While the common faults can be directly modelled it is not practical Lo assign states to all the other minor faults which might occur.

As discussed in [1] and [5], discriminative models directly model the posterior probabilities of the classes given the feature data and they assume that the classes are exhaustive. On the other hand, a *generative* model directly models the data likelihood $p(\underline{\theta}|\omega_i)$ and then determines poster ior class probabilities by application of Bayes' rule. Examples of generative classifiers include parametric models such as Gaussian classifiers and lllelrlory-based methods such as kernel density estimators and near neighbour models. Generative models are by nature well suited to novelty detection. However, there is a trade-offi because generative models typically arc doing more modelling than just scare.llillg for a decision boundary, they can be less efficient (than discriminant methods) in their use of the data. For example, generative models typically scale poorly with input dimensionality for fixed training sample size - see Dawid [(j] and Smyth [5] for further discussion.

4 Hybrid Models

A practical approach is Lo use both a generative and discriminative classifier arid add an extra "m -ı 1 th" state Lo the model to cover '(all other possible states" not accounted for by the known m states. Hence, the posterior estimates of the generative classifier are conditioned on whether or not the data is thought to come from one of the m known classes.

Let the symbol $\omega_{\{1,...,m\}}$ denot e the event that the true system state is one of the known states, and let $p(\omega_{m+1}|\underline{\theta})$ be the posterior probability that the data is from an unknown state. Hence, one can estimate the true posterior probability of individual known states as

$$\hat{p}(\omega_i|\underline{\theta}) = p_d(\omega_i|\underline{\theta}, \omega_{\{1,\dots,m\}}) \times (1 - p(\omega_{m+1}|\underline{\theta})), \qquad 1 \leq i \leq m \tag{1}$$

where $p_d(\omega_i|\underline{\theta}, \omega_{\{1,\dots,m\}})$ is the posterior probability estimate of state i as provided by a discriminative model.

The calculation of $p(\omega_{m+1}|\underline{\theta})$ can be obtained via the usual application Bayes' rule if $p(\underline{\theta}|\omega_{m+1})$, $p(\omega_{m+1})$, and $p(\underline{\theta}|\omega_{\{1,\dots,m\}})$ are known, i.e.,

$$p(\omega_{m+1}|\underline{\theta}) = \frac{p(\underline{\theta}|\omega_{m+1})p(\omega_{m+1})}{p(\underline{\theta}|\omega_{m+1})p(\omega_{m+1}) + p(\underline{\theta}|\omega_{\{1,\dots,m\}})\sum_i p(\omega_i)}. \tag{2}$$

In practice we have used non-informative Bayesian priors for $p(\underline{\theta}|\omega_{m+1})$ over a bounded space of feature values (details are available in a technical report [7]), although this choice of a prior density or data of unknown origin is basically ill-posed. The stronger the constraints which can be placed on the features, the narrower the prior density, and the better the ability of the overall model to detect novelty. If we only have very weak prior information, this will translate into a weaker criterion for accepting points which belong to the unknown category.

The term $p(\omega_{m+1})$ must be chosen based on the designer's prior belief of how often the system will be in an unknown state --- a practical choice is that the system is at least as likely to be in an unknown failure state as any of the known failure states.

The $p(\underline{\theta}|\omega_{\{1,\dots,m\}})$ term in Equation (2) is provided directly by the generative model. Typically this can be a mixture of Gaussians or a kernel density estimate over all of the training data (ignoring class labels). In practice, for simplicity of implementation we use a simple Gaussian mixture model. Furthermore, because of the afore-mentioned scaling problem with input dimensions, only a subset of relatively significant input features are used in the mixture model. A less heuristic approach to this aspect of the problem (with which we have not yet experimented) would be to use a method such as projection pursuit to project the data into lower dimensions and perform the input density estimation in this space. The main point is that the generative model need not necessarily work in the full dimensional space of the input features,

Integration of equation (1) into the hidden Markov model updating is straightforward and will not be derived — the model now has an extra state, "unknown." The choice of transition probabilities between the unknown and other states is once again a matter of design choice. For the antenna application at least, many of the unknown states are believed to be relatively brief transient phenomena which last perhaps no longer than a few seconds: hence the Markov matrix is designed to reflect these beliefs since the expected duration of any state $d[\omega_i]$ (in units of sampling intervals) must obey

$$d[\omega_i] = \frac{1}{1 - p_{ii}} \tag{3}$$

where p_{ii} is the self-transition probability of state ω_i.

5 Experimental Results

For comparison purposes we evaluated the results of 2 particular models. Each was applied to monitoring the servo pointing system of a DSN 34m antenna at Goldstone, California. The models were implemented within the LabView data acquisition package running in real-time on a Macintosh II at the antenna site. The models had previously been trained off-line on data collected some months earlier. 'There were 12 input features used. The experiment consisted of introducing

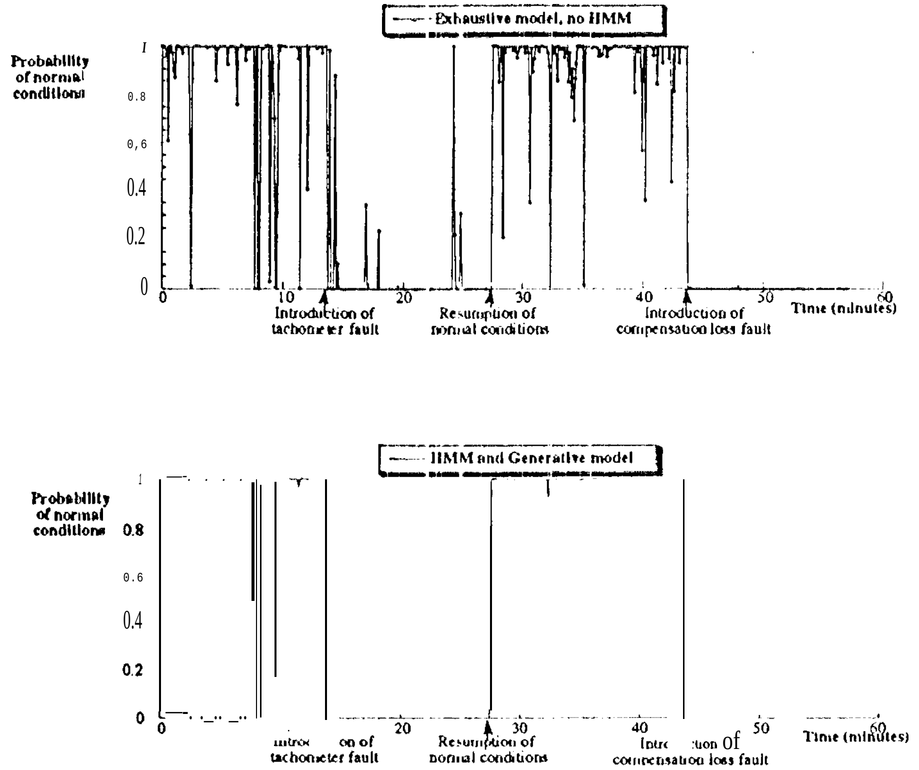

Figure 1: Estimated posterior probability of normal state (a) using no HMM and the exhaustive assumption (normal + 3 fault states), (b) using a HMM with a generative model (normal -t- 3 faults + other state).

hardware faults into the system in a controlled manner at 15 minutes and 45 minutes, each of 15 minutes duration.

Figure 1 (a) and (b) show each model's estimates over time that the system is in the normal state (space limitations precluded the inclusion of more detailed experimental results). Model (a) uses no HMM and assumes that the 4 known states are exhaustive - a single feedforward neural network with 8 hidden units was used as the discriminative model. Model (b) uses a H MM with 5 states, where a generative model (a Gaussian mixture model) and a flat prior (with bounds on the feature values) are used to determine the probability of the 5th state. The same neural network as in mode] (a) is used as a discriminator for the other 4 known states. The generative mixture model had 10 components and used only 2 of the 12 input features, the 2 which were judged to be the most senstive to system change. The parameters of the 11 MM were designed according to the guidelines described earlier, Known fault states were assumed to be equally likely with 1 hour MTBF's and with 1 hour mean duration. Unknown faults were assumed to have a 20 minute MTBF and a 10 second mean duration.

Model (a)'s estimates are quite noisy and contain a significant number of potential false alarms (highly undesirable in an operational environment). Model (b) is much more stable due to the smoothing effect of the HMM. Nonetheless, wc note that between the 8th and 10 minutes, there appear to be some possible false alarms: this data was classified into the unknown state (not shown). On later inspection it was found that large transients (of unknown origin) were in fact present in the original sensor data and that this was what the model had detected, confirming the result obtained independently by the classifier. It is worth pointing out that the model without a generative component (whether with or without the HMM) did in fact always detect a non-normal state at the same time, but incorrectly classified this state as one of the known fault states (these results are not shown).

6 Application, Issues

The ability to detect such previously unseen transient behaviour has important practical consequences: as well as being used to warn operators of servo problems in real-time, the model will also be used as a filter to a data logger to record interesting and anomalous servo data on a continuous basis. hence, potentially novel system characteristics can be recorded for correlation with other antenna-related events (such as maser problems, receiver lock drop during RF feedback tracking, etc.) for later analysis to uncover the true cause of the anomaly.

Based on these and related results, the basic model described here has recently been approved for inclusion as a functional requirement in the antenna controller design for all new DSN antennas. The first such antenna is currently being built at the Goldstone, California, DSN site and will become operational in 1994 -- similar antennas, also with onboard fault detectors of the type described here, will be constructed in Spain and Australia in the 1995-96 time-frame.

Acknowledgements

The research described in this paper was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and was supported in part by ARPA under grant number N00014-92-J-1860

References

1. P. Smyth and J. Mellstrom, 'Fault diagnosis of antenna pointing systems using hybrid neural networks and signal processing techniques,' in *Advances in Neural In formation Processing Systems 4*, J. E. Moody, S. J. Hanson, R. P. Lippmann (eds.), Morgan Kaufmann Publishers: San Mateo, CA, 1992, pp.667-674.

2. P. Smyth, 'Hidden Markov models for fault detection in dynamic systems,' submitted to *Pattern Recognition.*

3. M. D. Richard and R. P. Lipprnann, 'Neural network classifiers estimate Bayesian a posteriori probabilities,' *Neural* Computation, 3(4), pp.461--483, 1992.

4. J. Miller, R. Goodman, and P. Smyth, 'On loss functions which minimize to conditional expected values and posterior probabilities,' *IEEE Transactions on Information Theory*, *to* appear, July 1993.

5. P. Smyth, 'Probability density estimation and local basis function neural networks,' in *Computational Learning Theory and Natural Learning Systems*, T. Petsche, M. Kearns, S. Hanson, R. Rivest (eds.), Cambridge, MA: MIT Press, to appear, 1992.

6. A. P. Dawid, 'Properties of diagnostic data distributions,' *Biometrics*, 32, pp.647-658, Sept. 1976.

7. P. Smyth and J. Mellstrom, 'Failure detection in dynamic systems: model construction without fault training data,' *Telecommuncations and Data Acquisition Progress Report, vol. I* 1.?, Jet Propulsion Laboratory, Pasadena, CA, February 15th 1993.